様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２１ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０４１３３

研究課題名（和文）マルチエージェント深層学習による音声因子分解

研究課題名（英文）Speech factorization using multi-agent deep learning

研究代表者
篠田　浩一（Shinoda, Koichi）

東京工業大学・情報理工学院・教授

研究者番号：１０３４３０９７

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：音声に関する音声認識、音声合成、話者認識などの様々なタスクを担当するエージェントがお互いに競争・協調・調整しながら個々のタスクを学習する、マルチエージェントによる深層学習基盤を提供することを目的し、研究を行った。複数の音声を分離する音声分離において、雑音を明示的に扱い、それも分離する対象に含めることで、耐雑音性の高い音声分離を実現した。また、話者認識、音声認識の結果を用いて、話者特徴と音韻特徴を音声特徴から分離することにより、感情認識の性能を向上させることができた。

研究成果の学術的意義や社会的意義
音声には音韻性、話者性、感情、など様々な特徴が含まれているが、それらの特徴間の関係を陽にモデル化することにより、音声認識、話者認識、感情認識など様々なタスクの性能を向上させる方法論を提案し、その有効性を確認した。音声処理の多くの用途に応用が可能であり、すでに精神疾患の診断や、人間の性格の診断などに効果があることを確認している。また音声以外の画像など様々なメディアの処理においても有効であることが期待される。

研究成果の概要（英文）：We researched to provide a multi-agent deep learning infrastructure in which agents responsible for various tasks related to speech, such as speech recognition, speech synthesis, and speaker recognition, can learn individual tasks while competing, cooperating, and coordinating with each other. We achieved noise-tolerant speech separation by explicitly handling noise and including it as a separation target. In addition, using the results of speaker and speech recognition, we improved emotion recognition performance by separating speaker and phonological features from speech features.

研究分野：機械学習

キーワード：深層学習　音声認識　話者認識　話者分離　感情認識

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

音声情報処理では、深層学習を用いた技術革新が目覚ましい。例えば音声認識は現時点で理想的な環境下ならば人間と同等の性能をもつ。また、複数のタスク（例えば音声認識と話者認識）を一つのモデルで同時に学習するマルチタスク学習が効果をあげている。音声認識、音声合成、話者認識、言語識別、感情認識、音声強調(耐雑音音声認識)、音源分離、声質変換など、音声情報処理に含まれる分野は数多い。そして、これらは互いに密に関係している。例えば、話者の感情がわかれば、雑音下の音声認識率の向上に大きく寄与するであろう。これらすべてのタスクを対象としたマルチタスク学習を行えば、個々のタスクの性能は顕著に向上する可能性が高い。しかし、直接的な方法は実現が難しい。大量のデータにすべてのタスクのクラスラベルを付与することは難しいからである。

２．研究の目的

音声に関する音声認識、音声合成、話者認識などの様々なタスクを担当するエージェントがお互いに競争・協調・調整しながら個々のタスクを学習する、マルチエージェントによる深層学習基盤を提供する。エージェント群は、個々のタスクに関わる音声要素(因子)の間の含有・排他・共有などの関係を用いて音声データを因子分解することにより、個々のタスクの性能を高める。従来のマルチタスク学習に比べ、少量・非均一のデータでより高い性能を得ることを目的とする。研究期間内では、主に音韻、話者、雑音の三因子について開発評価を行う。結果として得られる因子の解析は音声科学の基礎研究に新たな知見を与える。また、画像・映像などの他のメディアに対し、新たなマルチタスク深層学習基盤を提供する。

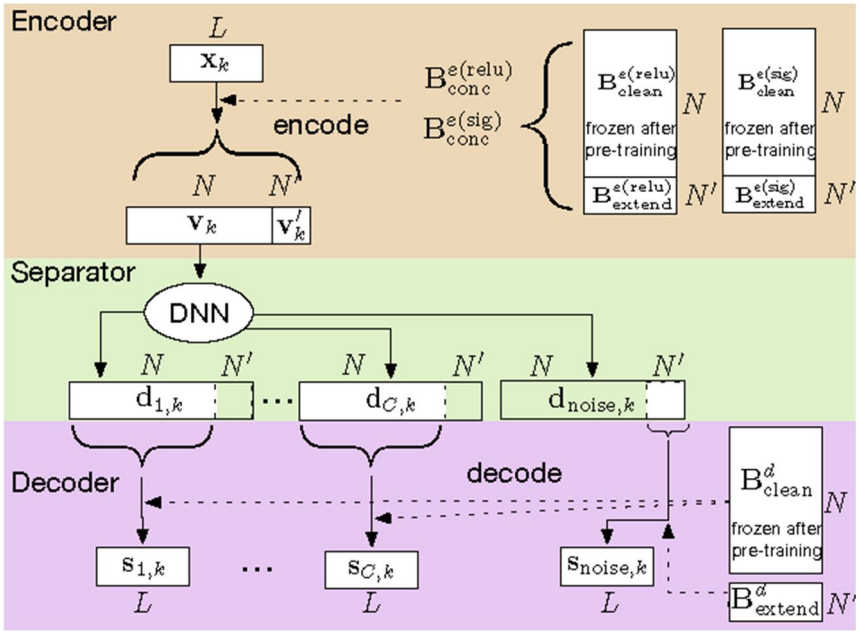
３．研究の方法

マルチタスク学習を行うためには、すべてのデータにラベルを付ける必要のない、半教師付き(semi-supervised)学習や弱教師付き(weakly-supervised)学習を用いる必要があるが、そのためには何らかの制約を新たに導入する必要がある。我々は以下の仮定を置いた。「音声は、ある特定のタスクのみに関する要素、あるタスクの組のみに関する要素、どのタスクにも関係する要素、どのタスクにも関係ない要素に分解できる。そして分解後に再構成することで元の音声を復元できる。」そして、これらの要素間の関係性を制約として用いるマルチタスク学習の方法論の確立を目指した。

４．研究成果

4.1 雑音に対して頑健な音声分離

深層学習を用いた音声分離が従来法よりも優れた性能をもち、盛んに研究されている。その中でも特に音声波形をそのままの形で入力し、畳み込みの処理により分離し、各々の話者の音声波形を出力する TasNet という手法が主流であった。しかしながら、この手法は、周囲雑音が大きく音声に雑音が含まれている場合にその分離性能が大きく低下する。そこで、我々は、畳み込み処理において、話者の基底に加え、雑音の基底を用いることで、分離性能を向上させる手法 TasNet with noise basis signals (TasNet-NB)を提案した。その構成図を以下に示す。



まず、符号化器(Encoder)により、音声波形から中間表現を得る。この際、従来の TasNet では、予め学習しておいた音声信号の基底と入力信号との畳み込み演算を行う。それに対し、TasNet-NB では音声信号の基底以外にノイズの基底も用意し、それを用いて雑音も陽に分離する。次に、分離器(Separator)では、中間表現を入力とし、そこから、分離された音声信号に対応する中間表現を出力する。最後に復号化器(Decoder)では、それぞれの中間表現から、音声に対応する部分とノイズに対応する部分を分けて取り出し、Encoder で用いたものと同じ基底を用いて、各々の音声信号とノイズの波形を生成し、出力する。

学習時には、まずノイズがない音声波形を入力し、ノイズ基底をゼロに固定して、音声基底を推定する。続いて、ノイズを加えた音声波形を入力し、音声基底を固定してノイズ基底を学習する。ここで、最初から大きいノイズを加えると、ノイズ基底の学習が不安定になるので、学習の初期段階では小さいノイズを加え、学習が進むにつれてノイズを大きくするカリキュラム学習を用いる。

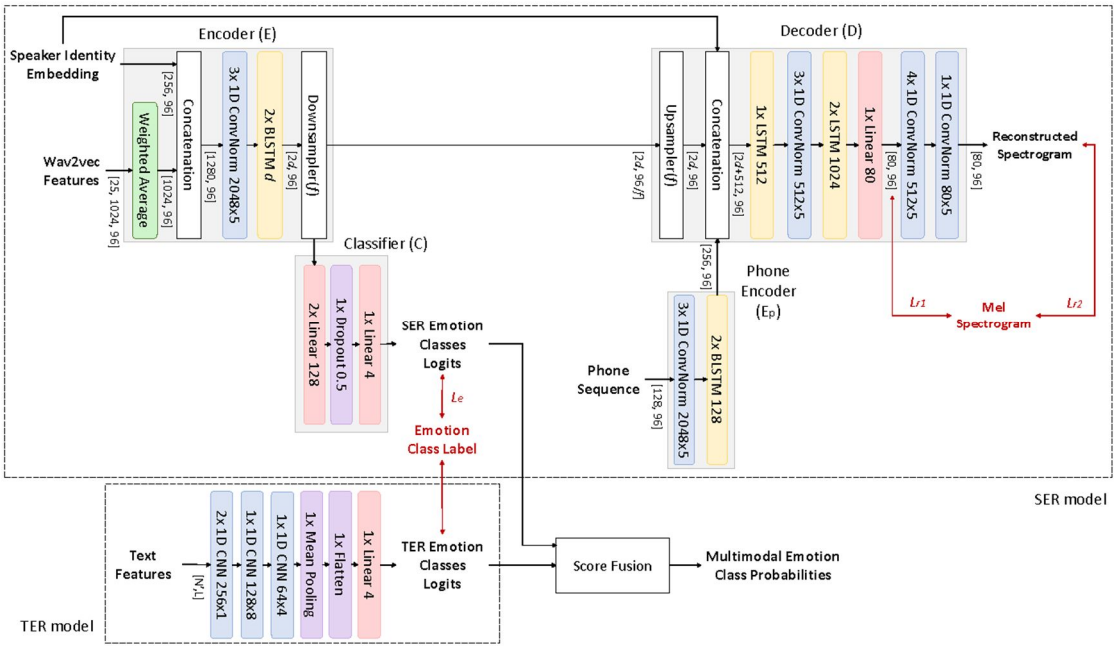
複数の読み上げ音声を事後処理で重畳し、それにノイズを加えたデータセットである、WHAM!データセットを用いて評価実験を行った。評価指標として、音声分離の指標 SI-SDR の分離前からの向上値 SI-SDRi を用いて評価したところ、従来の TasNet の値 13.7 に比べ提案した TasNet-NB は 14.6 と顕著に大きい値を示した。このことから提案手法の有効性を確認した。

この成果は IEEE の国際会議の一つであるアジア太平洋信号処理会議(APSIPA2021)に採択され、発表を行った。

4.2 話者性と音韻性の分離による感情認識

音声信号からの感情認識はマン・マシンインタフェースの重要な要素技術の一つである。この分野においても深層学習が有効であることが知られている。音声信号には、書きおこしに相当する音韻性を表す特徴や、誰がしゃべっているかを表す話者性を表すの特徴が支配的であり、感情を表す特徴は、それらの影響を大きく受けるため認識が難しい。そこで、我々は、音声の中間表現から、音韻性や話者性を示す特徴を

同定し、それを陽に取り除くことで、感情の特徴を抽出する方法を開発した。また、音声認識結果の書きおこしも併用するマルチモーダル認識により更に性能を向上させた。その構成図を以下に示す。



この手法では、自己符号化器(AutoEncoder)を用いて感情表現の獲得を行う。自己復号化器は符号化器(Encoder)と復号化器(Decoder)からなりこの　自己符号化器の入力は、自己教師付き学習により獲得された wav2vec 特徴であり、出力は音声スペクトルである。学習時には、入力音声のスペクトルと出力スペクトルとの間の再構成損失を最小にするように学習される。符号化により得られた中間表現には、音声の特徴が含まれていると考えられ、それを入力として感情の識別を行う。

しかし、このままでは、中間表現において、音韻特徴や話者特徴など、感情以外の特徴が支配的になり、感情特徴が十分に表現されていない。そこで、話者認識および音韻認識により得られた話者特徴と音韻特徴を、陽に符号化器および復号化器の両方に入力することにより、中間表現においてこれらの特徴がなくても元のスペクトルを再構成できるようにした。

さらに書きおこしを入力として、これも自己教師付き学習を行う、BERT などの Transformer ベースの特徴抽出器を用い、そこからも感情の特徴を抽出し、識別を行う。そして、前述の音声から得られた感情識別器が出力するスコアとの重みづけ和をとることで、感情ごとのスコアを得る。

代表的な感情認識のデータセットである IEMOCAP データセットを用い、感情 4 クラスの識別により提案手法を評価した。識別率は 70.1%となり、世界最高性能(発表当時)を達成した。音声認識に関する代表的なワークショップである、この成果は IEEE Automatic Speech Recognition and Understanding (ASRU) 2021 に採択され、発表を行なった。

3. 応用：うつ病の認識、性格認識

2)で述べた、感情認識方式は、識別器部分を修正するだけで、様々な音声特徴の認識に応用が可能である。我々はこれを、うつ病の認識や、性格診断に用い、良好な性能を得ている。例えば性格診断では、視覚的な特徴を用いる場合よりも、音声情報を用いるほうがより高い性能をもつことを示した。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 Makiuchi Mariana Rodrigues、Uto Kuniaki、Shinoda Koichi	４．巻 1
２．論文標題 Multimodal Emotion Recognition with High-Level Speech and Text Features	５．発行年 2021年
３．雑誌名 2021Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)	６．最初と最後の頁 350-357
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ASRU51503.2021.9688036	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kohei Ozamoto, Kuniaki Uto, Koji Iwano, Koichi Shinoda	４．巻 1
２．論文標題 Noise-Tolerant Time-Domain Speech Separation with Noise Bases	５．発行年 2021年
３．雑誌名 Proc. 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)	６．最初と最後の頁 624-629
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Wang Dongxiao、Kameoka Hirokazu、Shinoda Koichi	４．巻 1
２．論文標題 A Modified Algorithm for Multiple Input Spectrogram Inversion	５．発行年 2019年
３．雑誌名 Proc. ISCA Interspeech2019	６．最初と最後の頁 4569-4573
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/Interspeech.2019-3242	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Rodrigues Makiuchi Mariana、Warnita Tifani、Uto Kuniaki、Shinoda Koichi	４．巻 1
２．論文標題 Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection	５．発行年 2019年
３．雑誌名 Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop	６．最初と最後の頁 55-63
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3347320.3357694	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計8件（うち招待講演　2件／うち国際学会　3件）

１．発表者名
Nathania Nah, Takafumi Koshinaka,Koichi Shinoda

２．発表標題
Personality Recognition on Dyadic Interactions with Representation Learning

３．学会等名
電子情報通信学会SP IPSJ-SLP EA SIP　研究会

４．発表年
2023年

１．発表者名
Keisuke Ishikawa, Kuniaki Uto, Koji Iwano, Koichi Shinoda

２．発表標題
Team Takoyaki submission for VoxCeleb Speaker Recognition Challenge 2020

３．学会等名
The VoxSRC Workshop（国際学会）

４．発表年
2020年

１．発表者名
Kohei Ozamoto, Kuniaki Uto, Koji Iwano, Koichi Shinoda

２．発表標題
Noise-Tolerant Time-Domain Speech Separation with Noise Bases

３．学会等名
Proc. 13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）

４．発表年
2021年

１．発表者名
Koichi Shinoda

２．発表標題
Co-design of ML and HPC for video understanding

３．学会等名
1st International Workshop on Deep Video Understanding (DVU 2020)（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, Koichi Shinoda

２．発表標題
Speech-linguistic Multimodal Representation for Depression Severity Assessment

３．学会等名
情報処理学会研究報告, Vol.2019-SLP-130 No.8, Dec. 2019.

４．発表年
2019年

１．発表者名
篠田 浩一

２．発表標題
(基調講演) 深層学習と高性能計算

３．学会等名
xSIG2019, May 27, 2019

４．発表年
2019年

１．発表者名
尾座本 耕平, 岩野 公司, 宇都 有昭, 篠田 浩一

２．発表標題
雑音の基底信号を用いた 耐雑音性の高い時間領域音声分離

３．学会等名
信学技報, vol. 120, no. 399, pp. 63-67, Mar. 2021

４．発表年
2021年

１．発表者名
篠田 浩一

２．発表標題
巨大深層モデルの高速・省資源開発基盤とその応用

３．学会等名
情報処理学会 連続セミナー2021 第9回「AIトレンド：大規模モデルと生成モデル」, Oct. 2021（招待講演）

４．発表年
2021年

〔図書〕　計1件

1．著者名 日本音響学会、岩野 公司、河原 達也、篠田 浩一、伊藤 彰則、増村 亮、小川 哲司、駒谷 和範	4．発行年 2022年
2．出版社 コロナ社	5．総ページ数 208
3．書名 音声（下）	

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	井上　中順 (Inoue Nakamasa) (10733397)	東京工業大学・情報理工学院・助教 (12608)	
研究分担者	岩野　公司 (Koji Iwano) (90323823)	東京都市大学・メディア情報学部・教授 (32678)	
研究分担者	宇都　有昭 (Kuniaki Uto) (90345356)	東京工業大学・情報理工学院・助教 (12608)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関